**Weekly report**

날짜 : **2024-01-24**

연구원 : **최윤석**

* **이전 수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - CNE-join 결과로 나온 조인 가능한 테이블쌍 정보를 활용해 최대 n개의 테이블을 조인한 테이블을 벡터로 임베딩
      * 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리를 활용해 n-join-pair에 관한 insight를 얻기
        + n개의 테이블을 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리의 평균을 조인 결과의 순위를 나누는 기준으로 하여 성능평가를 진행
    - DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(use case diagram)
      * PPT로 작성
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - Table2Vec 논문을 참고하여 n-join-pair에 대한 아이디어 얻기
    - WITC 2024에 제출할 논문 작성하기
      * Full paper는 초록 버전 이후로 작성
* **수행결과**
  + **관계형 테이블 임베딩 :**
  + Table2Vec 논문을 참고하여 n-join-pair에 대한 아이디어 얻기
    - Table2Vec 논문의 주요 목적은 Table에 특화된 워드 임베딩 모델을 만드는 것. 테이블 검색 방법은 Ad Hoc Table 논문의 방법을 대부분 참고하였음
    - Table2Vec(Ad Hoc Table)에서 Query를 검색하였을 때 관련된 Table을 랭킹화해 사용자에게 제공하는 방법
      * Query에 포함된 단어집합에 대해 Word2Vec를 활용해 벡터집합으로 변환함.
      * Table에 포함된 단어집합(컬럼명, 레코드 값, 테이블 설명)을 Word2Vec를 활용해 벡터집합으로 변환.
      * Query벡터집합과 Table벡터집합간의 유사도 값을 계산함. 하나의 Query벡터집합과 모든 Table벡터집합간의 유사도 계산을 하여 유사도 값이 높은 순서대로 테이블을 랭킹화해 사용자에게 제시함. 유사도 계산은 총 4가지의 방법이 있음

텍스트, 도표, 라인, 평면도이(가) 표시된 사진

자동 생성된 설명

텍스트, 폰트, 스크린샷, 라인이(가) 표시된 사진

자동 생성된 설명  
 

* + - n-join-pair 아이디어
      * 기존의 n-join-pair 아이디어는 3개의 테이블을 조인한 테이블의 임베딩 벡터와 3개의 원천 테이블의 임베딩 벡터간의 거리 평균값이 작을수록 조인이 잘 되었다고 판단함. 테이블을 임베딩 벡터로 바꿀 때 윤종찬 연구원의 아이디어를 사용하였는데, 윤종찬 연구원의 아이디어 대신에 Table2Vec(Ad Hoc Table)의 아이디어를 변형한 뒤 사용하여 n-join-pair의 성능 평가를 진행
        + Table2Vec(Ad Hoc Table)의 아이디어는 Query로 Table을 찾는 것이었지만, 이를 Table로 Table을 찾는 방식으로 변형
    - Table2Vec 코드를 Github에서 다운받아 실행해보았지만 실패
      * Linux 환경에서 돌아가는 코드여서 윈도우에서 돌리기가 어려워 보임
      * Table2Vec 코드는 Table에 특화된 워드 임베딩 모델을 만드는 것이 목적인데, Table2Vec 논문의 실험에서 Google 뉴스 데이터로 사전 학습된 Word2Vec와 Table2Vec을 사용한 성능의 차이가 없어서 코드를 돌릴 필요가 없어 보임
      * 사전 학습된 Word2Vec 모델을 활용하여 Table2Vec(Ad Hoc Table)의 Table로 Table을 찾는 아이디어만 가지고 직접 코드를 만들어서 n-join-pair의 성능 평가를 진행할 예정
* **수행계획**
  + **관계형 테이블 임베딩 :**
    - CNE-join의 성능을 개선할 방법 고민
      * 조인하는 테이블의 개수를 3개 이상으로 확장해보기, 최적의 테이블 개수는 무엇인가? 이를 확인할 척도에 대해서도 고민
        + CNE-join의 결과로 나온 2개의 조인 가능한 테이블쌍 중에서 A-B, B-C 같이 테이블이 겹치는 쌍을 활용해 n-join-pair를 생성하는 알고리즘 및 이에 대한 융합도 산정 방안 고민 및 소스코드 작성(이전 CNE-join과의 일관성이 중요함)
      * 시간 단축 문제는 가장 나중으로 진행
    - Measuring and Predicting the Quality of a Join for Data Discovery 논문이 참고한 문헌을 읽고 n-join-pair 알고리즘 아이디어 얻기
    - CNE-join 결과로 나온 조인 가능한 테이블쌍 정보를 활용해 최대 n개의 테이블을 조인한 테이블을 벡터로 임베딩
      * 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리를 활용해 n-join-pair에 관한 insight를 얻기
        + n개의 테이블을 조인한 테이블의 임베딩 벡터와 원천 테이블의 임베딩 벡터간의 거리의 평균을 조인 결과의 순위를 나누는 기준으로 하여 성능평가를 진행
    - 연구원과 DataMap 개발(Type A, B)
    - 자동 융합 시나리오에 대한 명세를 작성해보기(use case diagram)
      * PPT로 작성
      * 융합을 할 때 Join을 할 지 Union을 할 지 확인하는 방법 – 나는 Join 파트
    - WITC 2024에 제출할 논문 작성하기
      * 저널 논문에 낸다고 생각하고 Full paper를 작성하고 WITC에는 그 일부를 제출
    - Table2Vec를 활용해 n-join-pair의 성능평가
* **기타사항**